# 基于 FasterR-CNN 的服务机器人物品识别研究 *

石　杰，周亚丽，张奇志

(北京信息科技大学 自动化学院，北京 100192)

**摘　要：**随着机器人在服务行业中的应用推广，尤其在家庭服务中有着重要的作用，对服务机器人的信息采集或目标识别需求也越来越强烈。传统的日用商品识别流程通常使用较为经典的图像识别和机器学习算法，如支持向量机(SVM)、随机森林或 Adaboost，然后利用目标图像的梯度、纹理或颜色的基本特征来对日用商品进行识别，可以在比较简单的背景中得到应用，但是在复杂的背景环境中很难有比较突出的表现，并且难以达到较高的准确率。目前在目标识别中表现比较优异的是卷积神经网络(CNN)，并成为很多目标识别场景中的首选。考虑到服务机器人的硬件配置成本，将基于区域的卷积神经网络(R-CNN)的快速算法 Faster R-CNN 引入系统中，并以 CPU 计算的方式进行物品识别。利用 CNN 网络提取图像特征，在其后面接入一个区域提议层。实验结果表明，将深度学习的识别方法应用到服务机器人平台是可行的，识别效果准确，且在实验中得到较好的检测效果。

**关键词：**服务机器人；深度学习；Faster R-CNN；物品识别

**中图分类号：**TP391.4　　doi: 10.3969/j.issn.1001-3695.2018.03.0311

## Item recongnition based on faster R-CNN in service robot

Shi Ji, Zhou Yali, Zhang Qizhi

(*School of Automation, Beijing Information Science & Technology University, Beijing* 100192, *China*)

**Abstract:** With the promotion and application of robots in the service industry, especially in the family service, the demand for information collection or target recognition for service robots is also getting stronger and stronger. Traditional commodity recognition processes typically use the more classic image recognition and machine learning algorithms such as support vector machines (SVM) , random forest or adaboost, then use the basic characteristics of the gradient, texture or color of the target image. It can be applied in a relatively simple background, but it is hard to have a more prominent performance in a complicated background environment, and it is difficult to achieve a high accuracy. At present, the convolution neural network (CNN) , which is superior in target recognition, has become the first choice in many target recognition scenarios. Considering the hardware configuration cost of service robot, Faster R-CNN, a fast algorithm of region-based convolutional neural network (R-CNN) , is introduced into the system and identified by CPU. The CNN network is used to extract image features and access to a regional proposal layer behind it. The experimental results show that it is feasible to apply the deep learning recognition method to the service robot platform. The recognition effect is accurate and the test results are good.

**Key words:** service robot; deep learning; Faster R-CNN; commodity recognition

## 0　引言

近年来，随着机器人行业的迅猛发展，服务型机器人开始进入大众视野，而人工智能相关技术的快速发展使得服务机器人得到了越来越多的技术支持。服务机器人具有非常广泛的应用前景，其发展是实现自动化服务的重要一步。同时，日用商品识别是智能服务行业的一个基础的研究问题，也是用于采集和分析超市日用商品大数据信息的前提和基础。在服务机器人平台实现物品识别是未来行业发展的重要趋势。一方面，服务机器人具有较为完善的硬件基础，可以完成基础的任务，功能强大；另一方面，深度学习在识别领域不仅能够完成多目标识别任务，还有很高的准确率。

物体检测识别依赖于计算机视觉的发展，一直以来是图像工程领域重要的研究热点。但由于技术发展的落后、大众对物体识别的认知度不强和算法应用场景的限制，物体检测真正快速的发展起始于 20 世纪 90 年代。人的肉眼通过视野内物体的

---

**收稿日期：**2018-03-20；**修回日期：**2018-05-16　**基金项目：**国家自然科学基金资助项目（11672044，11172047）；北京信息科技大学教改项目（2016JGYB09）；，2018 北京信息科技大学研究生科技创新项目

**作者简介：**石杰（1993-），男，河北保定人，硕士研究生，主要研究方向为深度学习图像处理；周亚丽（1968-），女，辽宁沈阳人，教授，博士，主要研究方向为机器人控制及信号处理究（zhouyali@bistu.edu.cn）.；张奇志（1963-），男，辽宁阜新人，教授，博士，主要研究方向为机器人控制.

纹理特征、颜色特征、深度信息等，定位、识别目标物体非常简单。然而计算机在处理图像时，面对的是数值化的三通道矩阵（即 RGB 矩阵），难以直接得到货架、物品这种抽象的概念，再加上图片背景的复杂度、物品摆放姿态的多样性和不同的光照强度等其他外界干扰因素，使得物体检测更加困难。与传统的物品识别检测相比，服务机器人采集的图像具有背景更为复杂、采集现场光亮强度不一、识别目标距离的远近、相同物品不同形状等难题。传统的物品识别方法主要采用基本图像处理方法，如背景建模[1]、HOG[2] (histogram of oriented gridients)、K-means 聚类算法[3]、特征点匹配算法[4]、加速稳健特征（speeded up robust features，SURF）[5]等。在传统的目标检测方法中，2001 年 Viola 等人在论文“鲁棒实时目标检测”中提出的 Viola-Jones 框架[6]得到了广泛的关注。这种方法速度快、相对简单，傻瓜相机的实时脸部检测就是使用这种算法，它的运算量很小。它使用 Harr 特征[7]来生成不同的简单的二分类，这些分类被级联的多尺度滑动窗口来处理，并且会及时丢弃错误分类。

深度学习作为机器学习领域的延伸已经众人皆知了，尤其在计算机视觉领域。与深度学习模型在图像分类领域优于传统模型类似，深度学习现在也是目标检测领域中最好的方法。过去几年深度学习目标检测方法有了很大的进步，纽约大学的研究人员在 2013 年提出了 Overfeat[8]，并在目标检测领域取得很大进展，他们提出了一种使用卷积的多尺度滑动窗口算法。在 Overfea 提出不久，伯克利大学的 Girshick 等人[9]提出了基于区域的卷积特征（region-based convolutional neural network，R-CNN）算法，这也是深度学习在目标检测领域的重大突破，该算法在目标检测效果上相比传统方法取得了 50%的性能提升。R-CNN 提出了一个三阶段的方法：a）使用区域提议方法提取可能目标，现在流行的方法是选择性搜索（selective search）[10]；b）使用卷积神经网络（convolution neural network ,CNN）在区域上提取特征；c）使用支持向量机（support vector machine,SVM）[11]对区域进行分类。虽然 R-CNN 在当时目标检测领域取得了非常不错的成绩，但是在训练过程中有很多问题。首先需要生成训练集的建议区域，然后在每个区域使用 CNN 特征提取器来提取特征，最后训练 SVM 分类器。这个过程需要消耗大量的时间。在 2015 年 Girshick[12]发表了 Fast R-CNN，这种方法迅速进化成一个纯深度学习方法。与 R-CNN 相似，它使用选择性搜索来生成区域建议；但与 R-CNN 不同的是，Fast R-CNN 在整张图上使用 CNN 来提取特征，然后在特征图上使用区域兴趣池化（region of interest，ROI），最后是一个反向传播网络来做分类和边框回归。这种方法不仅快，而且因为区域兴趣池化层和全连接层，该模型可以进行端对端的差分，训练也很容易。最大的不足是该模型仍旧依赖于选择性搜索，这也成为了模型推理阶段的一个瓶颈。

随后，Ren 等人[13]发表了 Faster R-CNN，这是 R-CNN 系列的第三个迭代。Faster R-CNN 增加了一个区域建议网络（region proposal network，RPN）[14]，试图摆脱选择性搜索算法，从而让模型实现完全端对端的训练。RPN 的作用是根据“物体”的分数来输出可能目标。这些目标区域被后面的 ROI 池化和全链接层来做分类。

随着深度学习在近些年的爆炸性发展和机器人在服务行业的广泛应用，本文在上述成果的基础上，将深度学习的方法应用到智能服务机器人平台。考虑到识别准确率和识别时间，本文主要利用 Faster R-CNN 作为目标检测算法，硬件平台采用自主研发的服务机器人。该机器人具有路径规划、行人跟踪、物体抓取、自主导航等功能。本文实现的是机器人自主识别物品，为未来实现自主抓取、导航打下基础。Faster R-CNN 主要有两种训练模式，一种是 2015 年 NIPS 中的“alternating optimization”（alt-opt）[15]方法，它的训练特点是迭代，先训练 RPN，随后用建议框去训练 Fast R-CNN，被 Fast R-CNN 微调的网络用来初始化 RPN，以此迭代；另一种是“End to End”，其训练特点是将 RPN 和 Fast R-CNN 融合到一个网络中进行训练，在每次随机梯度下降（SGD）迭代过程中，前向传递时 RPN 产生区域建议框，这些建议框被当做固定的、提前计算好的建议框来训练 Fast R-CNN 检测器。反向传递时，对于共享层来说，来自 RPN 的损失函数与 Fast R-CNN 的损失函数结合。但是这种方法不考虑边界框（bounding boxes），忽略了建议框的坐标也是网络的输出。这两种算法主要推荐第二种方法，因为“End to End”使用的显存小，而且训练更快，同时准确率略高于 alt-opt 算法，实验也将会对比两种算法的测试效果。本文主要将深度学习 Faster R-CNN 应用到家庭服务机器人平台，通过竞赛和实际实验得出了将深度学习算法应用服务机器人的可行性和有效性，并且识别效果要强于传统方法。

## 1　物品识别算法

### 1.1　Fast R-CNN 算法

虽然 R-CNN 使用了选择性搜索等预处理步骤来提取潜在的 bounding box 作为输入，但是 R-CNN 仍会有严重的速度瓶颈。原因也很明显，就是计算机对所有区域进行特征提取时会有重复计算。Fast R-CNN 正是为了解决这个问题诞生的。

Fast R-CNN 算法解决了 R-CNN 算法的三个问题：

a)测试速度慢。Fast R-CNN 解决方法：Fast R-CNN 将整张图像归一化后直接送入 CNN。在最后的卷积层输出的特征图上，加入建议框信息，使得在此之前的 CNN 运算得以共享。

b)训练速度慢。 Fast R-CNN 解决方法：Fast R-CNN 在训练时，只需要将一张图像送入网络，每张图像一次性提取 CNN 特征和建议区域，训练数据在 GPU 内存里直接进 Loss 层，这样候选区域的前几层特征不需要再重复计算，并且不再需要把大量数据存储在硬盘上。

c)训练所需空间大。Fast R-CNN 解决方法：Fast R-CNN 把类别判断和位置回归统一用深度网络实现，不再需要额外存储。

Fast R-CNN 模型同样采用 CNN 的结构。图 1 为 CNN 的传统架构。采样层交替插入在卷积层中，这样图像在经过卷积层

后所提取的特征，再经过筛选组合形成新的特征图，这个特征图是对原始图片更抽象的描述，最后把这些更加抽象的参数归一化成计算更加方便的一维数组，就形成全链接层，通过得分函数进行物品的分类、检测。

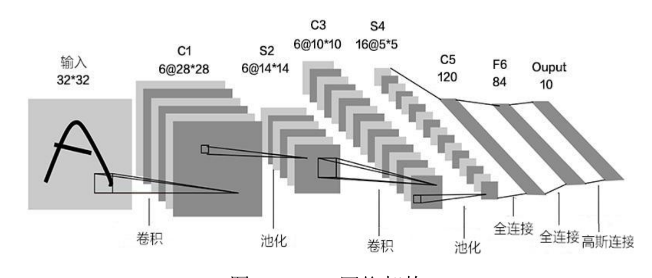



图 1　CNN 网络架构

## 1.2　Faster R-CNN 算法

Faster R-CNN 的一大特点是利用 RPN，其经过训练可以直接预测出建议框，比选择性搜索预测所提取的预测框数量更少、速度更快，且 RPN 的预测绝大部分在 GPU 中完成，同时卷积网和 Fast R-CNN 部分共享，这些对于提升物品检测速度起到至关重要的作用。

Faster R-CNN 不同于其他分类检测网络的两个关键点是：a)使用区域推荐网络替代原有的选择性搜索方法产生建议窗口；b)产生建议窗口的卷积神经网络和目标检测的卷积神经网络的共享。Faster R-CNN 的整体框架大致为：

•Faster R-CNN 把整张图片输入 CNN，进行特征提取

•生成区域推荐窗口，Faster R-CNN 用区域推荐窗口生成建议窗口，对于输入的每一张图片都会生成 300 个建议窗口；

•Faster R-CNN 把建议窗口映射到 ROI 生成的最后一层特征图上；

•利用 Softmax Loss 和 Smooth L1 Loss 对分类概率和边框回归（bounding box regression）联合训练。

图 2 是 Faster R-CNN 网络结构。

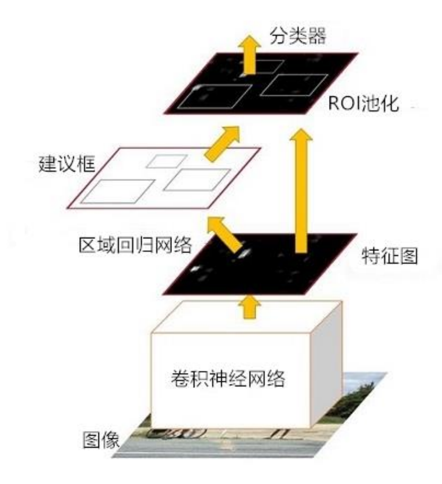



图 2　Faster R-CNN 网络结构

### 1.2.1 区域推荐网络（RPN）

为了将一个物体可以在不同尺寸下识别出来，有两种主要方式：a）对输入进行裁剪；b）对特征图进行不同大小的滑框卷积运算。而区域推荐网络采取了不同的方式。RPN 的输入是任意大小的图像，输出是一组打过分的候选框。

RPN 的核心思想是使用卷积神经网络直接产生建议区域，使用的方法本质就是滑动窗口，并且只需要在最后的卷积层上滑动一遍，在每一个滑动窗口的位置，同时预测 k 个区域推荐，所以回归层有 4k 个输出，k 个 box 的坐标编码；分类层输出 2k 个得分。生成训练数据的过程为：先检查 anchor 覆盖目标的真正区域（ground truth）是否超过 75%，如查超过就将当前 anchor 的目标分类标记为“存在”；如果没有超过就选择一个覆盖比例最大的标记为“存在”，即对每个建议框是目标/非目标的估计概率。因为 anchor 机制和边框回归可以得到多尺度长宽比的 RPN 网络也是全卷积网络（fully-convolutional network，FCN[16]），可以针对生成检测建议框的任务端对端地训练，能够预测出物体的边界和分数，只是在 CNN 上额外增加了 2 个卷积层（全卷积层 cls 和 reg）。一般为了应对尺寸迥异的物体，Faster R-CNN 应用了 3 种长宽比类型不同的 Anchor，即 1：1、2：1、1：2。再将这 3 种 Anchor 分别用 3 个尺度缩放，即 128、256、512，共 9 种类型的 Anchor boxes。这 9 种窗口在卷积特征图上经过卷积运算形成 256 维向量（Faster R-CNN 有三种训练模式，即 ZF[17]、VGG[18]、VGG16。这里的 256 维针对的是 ZF 模型，VGG 模型需要形成 512 维的向量），最后挑选出得分最高的 300 个窗口作为最终的建议窗口。RPN 模型网络结构如图 3 所示。

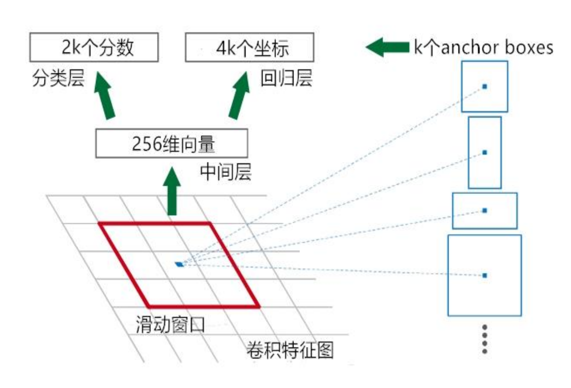



图 3　RPN 模型网络结构

RPN 的目标函数是分类和回归损失的和。根据文献[19]分类采用了交叉熵，回归采用了稳定的 Smooth $L1$，公式为

$$\text{Smooth}_{L1}(x)=\begin{cases}0.5x^2 & |x|<1\\ |x|-0.5 & \text{others}\end{cases} \tag{1}$$

整体的损失函数具体为

$$L(\{p_i\},\{t_i\})=\frac{1}{N_{cls}}\sum_i L_{cls}(p_i,p_i^*)+\lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i,t_i^*) \tag{2}$$

在式(2)中，$i$ 是一个 mini-batch 中 anchor 的索引；$p_i$ 是 $i$ 作为一个目标的 anchor 的预测概率。如果这个 anchor 为正，则 ground-truth 的标签 $p_i^*$ 就为 1；如果这个 anchor 为负，则 $p_i^*$ 就为 0。$t_i$ 表示预测边界框的 4 个参数化坐标的矢量，并且 $t_i^*$ 是与 anchor 相关联的 ground-truth 的矢量。

RPN 中分类损失函数为

$$L_{reg}\left(t_i,t_i^*\right)=-\log\left[p_i^* p_i+\left(1-p_i^*\right)\left(1-p_i\right)\right] \tag{3}$$

对于边界回归框，本文采用以下 4 个坐标的参数公式：

$$\begin{aligned} &t_x=\left(x-x_a\right)/w_a,t_y=\left(y-y_a\right)/h_a,\\ &t_w=\log\left(w/w_a\right),t_h=\log\left(h/h_a\right),\\ &t_x^*=\left(x^*-x_a\right)/w_a,t_y^*=\left(y^*-y_a\right)/h_a,\\ &t_w^*=\log\left(w^*/w_a\right),t_h^*=\log\left(h^*/h_a\right), \end{aligned} \tag{4}$$

其中：$x$、y、w 和 h 表示框的中心坐标及其宽度和高度；变量 $x$、$x_a$、$x^*$ 分别代表预测框、anchor 框和 ground-truth（同样用于 y、w、h）。这可以被认为是从一个 anchor 框到附近的 ground-truth 的边界回归框。

1.2.2 训练 RPN 网络

RPN 网络一般通过方向传播和随机梯度下降（stochastic gradient descent，SGD）方法来进行端对端训练。本文遵循“以图像为中心”的采样策略来训练这个网络。每个 mini-batch 包含很多正、负样本 anchor 的单个图像。对所有 anchor 的损失函数进行优化是可能的，但是使优化结果偏向于负样本，因为它们是占据主导地位的。相反，本文在图像中随机采样 256 个 anchor，以计算 mini-batch 的损失函数，取样的正负样本比例高达 1:1。如果图像中的正样本少于 128 个，则使用负样本来填充 mini-batch。

## 2　实验研究

### 2.1　实验平台

图 4 是北京信息科技大学家庭服务机器人团队的 Sun@Home 机器人。



图 4　Sun@Home 家庭服务机器人

本实验平台主要由一个 kinect Ⅱ代摄像头[20]、可调升降机构、3 自由度机械臂、全向轮底盘和一个检测范围为 270° 的激光雷达组成。本次实验使用到的传感器装置是 kinect Ⅱ代代摄像头。

相比于第一代 kinect，第二代 kinect 感应器具备了更高的分辨率和色彩识别度，使识别更加精准。其彩色摄像头分辨率为 1920*1080；深度感知器分辨率为 512*424；帧率为 30 fps；检测范围为 0.5~4.5 m。同时其拍摄的高清图片可提高算法对物品定位、识别的精度和准确度。

### 2.2　数据集

由于 Faster R-CNN 官方使用的是 1 000 类 ImageNet 数据集，在分类检测时有比较高的准确率和较好的识别效果，所以本次实验使用官方的预训练模型对重新制作的 VOC 数据集进行微调，这样在不用采集千万张图片基础上也可以得到较为稳定的参数。

1）数据集的采集与制作　由于机器人自身带有 kinect 传感器，所以使用 kinect 对 10 类目标物品进行录制，并分解成 2 010 张 JPG 图片，Labels 共有 2 010 个 XML 文件。训练集和训练验证集共有总图片的 90%，即 1 809 张；测试集和验证集共有总图片的 10%，即 201 张。最后将 XML 文件，训练集、测试集、验证集，图片分别放入 Annotation、ImageSets 下的 Main、JPEGImages 中，最终制作成 VOC2007 数据集。图 5 为机器人采集数据的实际场景。

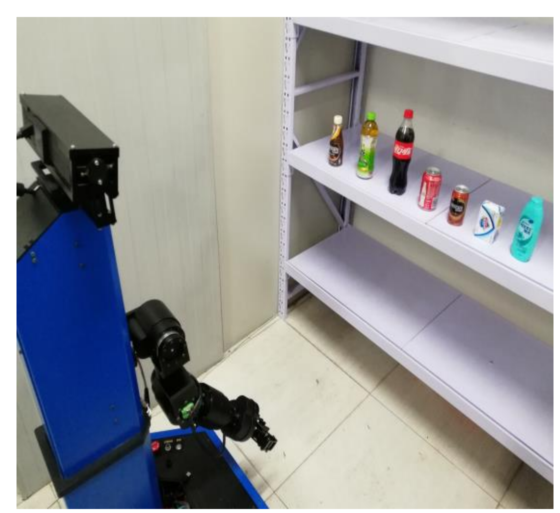


图 5　机器人采集数据的实际场景

2）数据集的具体内容　可乐（cola）、牛奶（milk）、酸奶（yoghourt）、牙膏（toothpaste）、咖啡（coffee）、绿茶（tea）、沐浴露（bath）、洗发水（shampoo）、水（water）、香皂（soap）。

3）数据集的训练平台　考虑到服务机器人的硬件设计没有加入高性能 GPU，训练使用的服务器配置是：因特尔酷睿 i7-6700K CPU 4 000 GHz*8，Geforce GTX 1080，ubuntu14.04 64-bit，硬盘 SSD 256 GB。同时训练不同的模型所用的时间也不同。

### 2.3　提取特征图

Faster R-CNN 特征提取的核心是 CNN，它会学习到物品的颜色、形状、纹理等特征，同时也学习到背景，利用 Caffe 的可视化方法可以直观地看到目标的关键信息。图 6 是各卷积层提取的特征图。其中 conv1 和 conv2 对浅层特征的提取，如物品的颜色、边缘等；conv3 提取到目标的纹理特征；conv4、conv5 提取到更为关键的特征。而图 7 是池化层对特征的精提取。从

图中可以看出物品的关键信息更加明显。

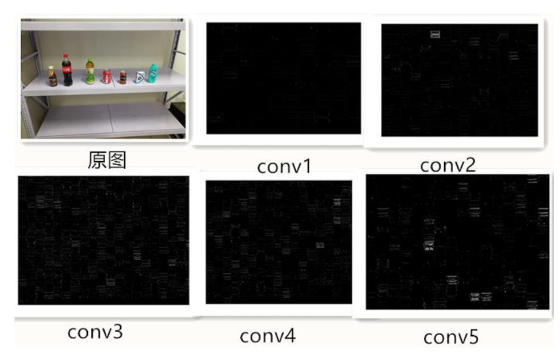



图 6　各卷积层提取特征图

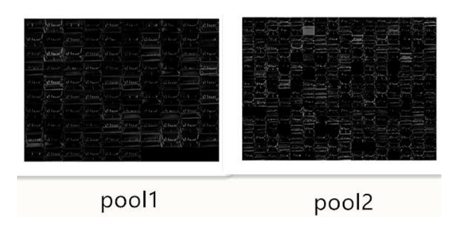



图 7　池化层对特征的精提取

## 2.4　实验结果与分析

本实验中误检测为检测到物品但错误地识别；漏检测为未检测到物品。

1）Faster R-CNN 与 Fast R-CNN 实验对比

在本实验中采用了 Faster R-CNN 算法，并与 Fast R-CNN 的实验效果进行了对比。由于 Faster R-CNN 多训练了 RPN 网络，所以在识别准确率上有很大的不同。对于 RPN 网络的训练，同样采用预训练模型对 RPN 微调的方法，将 RPN 训练得到的候选框训练 Fast R-CNN 网络；将 Fast R-CNN 网络训练得到的参数固定卷积层微调 RPN；随后在固定 Fast R-CNN 卷积层的前提下，通过 RPN 训练得到的候选框对 Fast R-CNN 进行微调；重复上述步骤，直到网络收敛。

测试实验图中物品摆放的原则是:尽可能将外形、颜色相近的物品摆放在一起；同品牌的物品摆放间隔尽可能大。同时物品间的距离对最终的测试结果有一定影响。

图 8 是实际测试场景图，包含了所有数据集中的物品。测试结果预期目标：正确地定位识别出所有物品。图 9 是 Fast R-CNN 识别效果。图 10 是 Faster R-CNN 识别效果。

从图中可以看出，将两种算法分别应用在机器人平台中得到的检测结果大不相同。由图 10 可见，机器人使用 Fast R-CNN 算法，正确地识别出咖啡、绿茶、沐浴露、牛奶、可乐、洗发水，但将酸奶误识别成牛奶和未识别出香皂，这可能是由于酸奶和奶有相似的颜色和外形特征，香皂与背景的颜色相近并且目标较小，在提取其特征时会得到不明显的特征，所以造成未识别。而由图 11 可看到，机器人使用 Faster R-CNN 算法可以正确地定位、识别出所有物品。Fast R-CNN 网络对于物品识别的效果以及准确率都没有 Faster R-CNN 网络好，Fast R-CNN 网络出现误检测和漏检测，而 Faster R-CNN 网络将物品全部识别准确。



图 8　测试场景图




图 9　Fast R-CNN 识别效果




图 10　Faster R-CNN 识别效果

再加载其他测试图继续进行误检、漏检实验。实验物品主要包括奶茶、一次性纸杯、香皂、纸卷，其中只有香皂是数据集中的物品，奶茶、一次性纸杯、纸卷不是数据集中物品。测试结果预期目标：只定位识别出香皂，其他物品不识别。图 11 为 Fast R-CNN 测试图。图 12 为 Faster R-CNN 测试图。




图 11　Fast R-CNN 测试图




图 12　Faster R-CNN 测试图

选取该四件物品作为误检、漏检测试的原因是：奶茶、一次性纸杯、纸卷的形状与颜色和训练数据集中的物品，像奶、酸奶、香皂等，有很高的相似度。

对比图 11 和 12 的实验结果可以看出，Fast R-CNN 能够正

确地识别出香皂，但将纸卷同样识别成香皂，并且置信度（即准确度）只有 74.2%，造成这个结果的原因是 Fast R-CNN 在测试时提取到纸卷和香皂具有相同的颜色、纹理特征。而 Faster R-CNN 可以正确地识别出香皂，同时不会造成误检。两种算法都不会误检奶茶、一次性纸杯。

下一步选取多件物品，其中只有一件不是训练数据集中的物品，同时实验测试角度，与数据集中部分物品采集样本的角度有些差别，即实验两种算法的环境适应性。实验物品包括香皂、沐浴露、咖啡（两种）、奶茶、洗发水，其中除奶茶外，剩余物品全部是数据集中的物品。测试结果预期目标：除奶茶外，正确地定位识别所有物品。图 13 为 Fast R-CNN 测试图。图 14 为 Faster R-CNN 测试图。




图 13　Fast R-CNN 测试图




图 14　Faster R-CNN 测试图

对比图 13 和 14 的实验结果可以发现，两种算法都能够正确地识别出部分训练数据集中的物品，但将罐装咖啡误识别成可乐。经过多次实验发现，在该角度下两种算法提取到的罐装咖啡与可乐有部分相似的特征，数据集中罐装可乐没有该角度下的样本，而可乐有多角度的样本，同时在该光照下的罐装咖啡的颜色特征与可乐的颜色特征相似，因此造成此类识别结果。所以，机器人控制 kinect 拍摄图片的角度与数据集中的物品拍摄角度应高度相似，即对样本进行尺度变换和增强，这对算法在识别过程中提取非常突出的物品特征和对物品的定位、识别有重要的作用。

最后改变实验图片采集的角度，分别使用 Fast R-CNN 和 Faster R-CNN 两种算法对多张不同角度的图片进行测试。通过大量实验结果可以得出，Fast R-CNN 和 Faster R-CNN 测试实验结果都并不非常理想。Fast R-CNN 能够正确地识别数据集中的部分物品；而 Faster R-CNN 除了能够正确地识别出数据集中的物品外，对非数据集中的物品产生了误检。造成上述两个实验结果的原因在于机器人拍摄物品的角度与训练数据集中物品拍摄角度不同，所以可适当地增加训练数据，对图片进行尺度变换，通过数据增强的方式，弥补机器人平台采集数据的劣势。

最后结合上述实验结果和其他测试图片的实验结果，将 Fast R-CNN 和 Faster R-CNN 网络在家庭服务机器人竞赛中的各项参数统计在表 1 中。

表 1　Fast R-CNN、Faster R-CNN 参数对比

| 算法 | mAP | 训练耗时 | 测试耗时 | 误检率 | 漏检率 |
|---|---|---|---|---|---|
| Fast R-CNN | 75% | 约 10h | 2.1s | 20% | 5% |
| Faster R-CNN | 90% | 约 14h | 1.5s | 6% | 4% |

由表 1 分析可得，Fast R-CNN 检测准确率在 75%左右，训练耗时约 10 h，测试耗时 2.1 s 左右，误检率 20%，漏检率 5%；Faster R-CNN 检测准确率在 90%左右，训练耗时约 14 h，测试耗时 1.5 s 左右，误检率 6%，漏检率 4%。由于训练方式的不同，Faster R-CNN 多训练一个 RPN 网络，所 以训练耗时要比 Fast R-CNN 略长。考虑到开发成本较高和服务机器人的内部空间不足等问题，虽然在耗时上两者都没有达到实时的级别，但 Faster R-CNN 的测试耗时明显低于 Fast R-CNN。分析原因是 Faster R-CNN 使用卷积网络自行产生建议框，并且和目标检测网络共享卷积网络，使得建议框数目从原有的约 2 000 个减少为 300 个，且建议框的质量也有很大的提高。

2）End to End 训练方法与 alt-opt 训练方法对比

Faster R-CNN 提供了两种不同的训练算法，两种方法都会得到优于 Fast R-CNN 的检测效果。

首先对多目标进行实验。机器人拍摄到角度正常的测试图片，为了增加实验难度，选择的实验物品有相似的形状特征、颜色特征，且会出现非数据集中的物品。

通过多次实验可以得出以下结论：在服务机器人平台拍摄的测试图片角度正常的情况下，机器人使用 Faster R-CNN 中的 alt-opt 算法检测的准确率接近 End to End 检测的准确率，在对正常角度图片测试时，End to End 算法可以正确地定位识别出数据集中的物品，同时不会检测非数据集物品。

为了进一步检测算法效果，下一步将测试非正常角度的物品。实验物品包括洗发水、洁面乳、瓶装咖啡。三种物品有相似的轮廓，但在两种算法中得出了不同的结果，其中洗发水和瓶装咖啡是数据集中的物品，洁面乳不是数据集中的物品。测试结果预期目标：可以正确地定位识别出洗发水和瓶装咖啡，不会识别洁面乳。图 15 为 alt-opt 训练方法测试图。图 16 为 End to End 训练方法测试图。

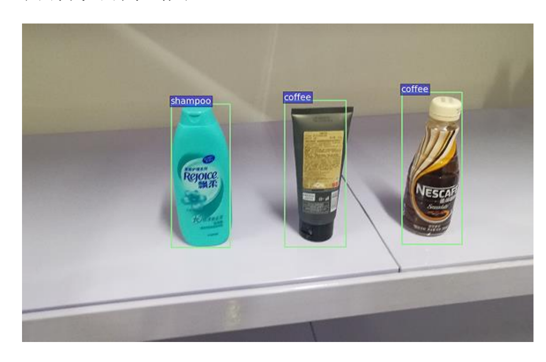



图 15　alt-opt 训练方法测试图

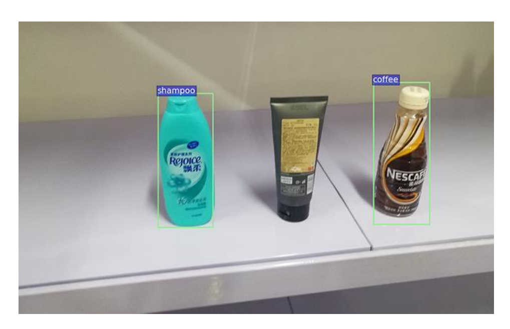


图 16　End to End 训练方法测试图

通过多次实验可以看出，图 15 使用 alt-opt 训练方法得到的模型正确地将洗发水和瓶装咖啡识别出来，但出现了误检测，洁面乳被识别成咖啡。出现这种结果的原因很可能是洁面乳和瓶装咖啡有相似的暗色和轮廓，在特征提取时得到了近似的特征。而图 16 是使用 End to End 方法测试的结果，该方法训练的模型正确地检测到洗发水和瓶装咖啡，没有出现误检测。同时 End to End 算法检测时间较 alt-op 算法检测时间略快，这是由于 alt-opt 算法在训练模型过程中分四个阶段，End to End 算法不分阶段，所以 alt-opt 算法较 End to End 算法产生更多的权重值。综上所述，对于本实验平台，End to End 方法更适用。

综合两种训练算法，表 2 为多次实验测试结果的参数。

表 2　alt-opt 与 End to End 参数对比

| 算法 | mAP | 训练耗时 | 测试耗时 | 误检率 | 漏检率 |
|---|---|---|---|---|---|
| alt-opt | 89% | 约 14h | 1.6s | 6% | 5% |
| End to End | 90% | 约 11h | 1.5s | 5% | 4% |

由表 2 分析可得，alt-opt 检测准确率在 89%左右，训练耗时约 14 h，测试耗时 1.6 s 左右，误检率 6%，漏检率 5%；End to End 检测准确率在 90%左右，训练耗时约 11 h，测试耗时 1.5 s 左右，误检率 5%，漏检率 4%。由于 alt-opt 算法训练分为四个阶段，每个阶段在训练完后都会进行模型检测，所以在训练过程中耗时更多。在检测速度、误检率和漏检率方面，End to End 训练方法略胜一筹。而 End to End 算法在测试时占用的内存较 alt-opt 算法小，因此检测速度更快。

分析实验结果：目标检测网络中，在特定场景下，Faster R-CNN 的网络检测性能要优于 Fast R-CNN；对于本实验平台应用的 Faster R-CNN 算法中，End to End 训练方法要略优于 alt-opt 算法。在实际测试中，机器人使用 Faster R-CNN 能够识别出更多的物品。同时，适当地增加样本集中物品的数量会对最终检测结果有更好影响，多卷积层可提取更多的物品特征，保证识别准确率。

## 3　结束语

近年来随着人工智能技术的飞速发展，传统的物品识别方法不仅效率低、检测速度慢，而且误检率和漏检率高。然而深度学习的方法将物品特征交给神经网络去提取，通过反向传播算法与正向传播构成反馈自动地调整学习到的参数，避免了单个算法在识别过程中的劣势，很大程度上解决了这些问题。本文利用深度学习中识别效果较好的 Faster R-CNN 算法，将物品识别方法转向将深度学习应用到家庭服务机器人平台，通过选择适用于实验平台的训练算法实现机器人自主识别物品并提高准确率。通过实验和竞赛结果对比发现，Faster R-CNN 算法应用在服务机器人可达到 90%的识别准确率。考虑到服务机器人平台的造价和内部空间的前提下，本次实验未配置 GPU，因此所有实验结果均是在 CPU 上得到，在 CPU 上运行 End to End 训练方法的检测速率在 1.5 s 左右，运行 alt-opt 训练方法的检测速率在 1.6 s 左右，准确率均在 90%左右，相比传统图像识别算法，深度学习更胜一筹。因此采用 Faster R-CNN 加 End to End 训练算法更适用于服务机器人平台。但是该方法对于光线有一定的要求，如果光线过暗，机器人拍摄的图片可能无法进行正常的识别。经过多次实验发现，较小目标的识别也有一定困难。考虑到 GPU 对算法运行的速率和准确性，团队最终会依据硬件平台加入适用版本的 GPU。同时，实时性也是工业级机器人发展的方向。对于这些问题，Sun@Home 物品识别团队将会继续开发和深入研究。

## 参考文献：

[1] 宋欢欢. 复杂场景下背景建模方法的研究与实现 [D]. 南昌: 南昌大学, 2015. (Song Huanhuan. Reseearch and implementation of background modeling method under complex scene [D]. Nanchang: Nanchang University, 2015. )

[2] Taigman Y, Yang Ming, Ranzato M A, *et al*. Deepface: closing the gap to human-level performance in face verification [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S. l. ] : IEEE Press, 2014: 1701-1708.

[3] 潘威, 左欣, 沈构强, 等. 物品识别系统的设计与实现 [J]. 科技视界, 2015 (5): 167. (Pan Wei, Zuo Xin, Shen Gouqiang, *et al*. The design and implementation of item identification system [J]. Science & Technology Vision, 2015 (5): 167. )

[4] 林汀. 图像特征点匹配算法的研究 [J]. 现代计算机: 专业版, 2016 (10): 28-32. (Lin Ting. Research on feature points matching algorithm sequence image [J]. Modern Computer, 2016 (10): 28-32. )

[5] 胡修祥. 基于 RGB-D 数据的物品识别与定位 [D]. 天津: 中国民航大学, 2016. (Hu Xiuxiang. Objeect recongnition and localization based on RGB-D data [D]. Tianjin: Civil Aviation University of China, 2016. )

[6] Viola P, Jones M J. Robust real-time face detection [J]. International Journal of Computer Vision, 2004, 57 (2): 137-154.

[7] 史东承, 倪康. 压缩感知 W-HOG 特征的运动手势跟踪 [J]. 智能系统学报, 2016, 11 (1): 124-128. (Shi Dongcheng, Ni Kang. Motion gesture tracking based on compressed sensing W-HOG features [J]. CAAI Trans on Intelligent Systems, 2016, 11 (1): 124-128. )

[8] Sermanet P, Eigen D, Zhang X, *et al*. OverFeat: integrated recognition, localization and detection using convolutional networks. [C]// Advances in Neural Information Processing Systems. [S. l. ] : ICLR Press, 2014: 1055-1061.

[9] Girshick R, Donahue J, Darrell T, *et al.* Rich feature hierarchies for accurate object detection and semantic segmentation [C]// Proc of ImageNet Large-Scale Visual Recognition Challenge Workshop. [S. l. ] : ICCV Press, 2013: 10-15.

[10] Girshick R, Donahue J, Darrell T, *et al.* Rich feature hierarchies for accurate object detection and semantic segmentation [C]// Proc of ImageNet Large-Scale Visual Recognition Challenge Workshop. [S. l. ] : ICCV Press, 2013: 10-15.

[11] Kazemi F M, Samadi S, Poorreza H R, *et al.* Vehicle recognition using curvelet transform and SVM [C]// Proc of the 4th International Conference on Information Technology. [S. l. ] : IEEE Press, 2007: 516-521.

[12] Girshick R. Fast R-CNN [C]// Proc of IEEE International Conference on Computer Vision. [S. l. ] : ICCV Press, 2015: 10-15.

[13] Ren S, He K, Girshick R, *et al.* Faster R-CNN: towards real-time object detectionwith region proposal networks [C]// Proc of Conference on Neural Information Processing Systems. [S. l. ] : NIPS Press, 2015: 1-15.

[14] Felzenszwalb P F, Girshick R B, McAllester D, *et al.* Object detection with discriminatively trained part-based models [C]// Proc of IEEE Transactions on Pattern Analysis and Machine Intelligence. [S. l. ] : TPAMI Press, 2010: 201-205.

[15] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[16] Zeiler M D, Fergus R. Visualizing and understanding convolutional neural networks [C]// Proc of European Conference on Computer Vision. 2014.

[17] 万士宁. 基于卷积神经网络的人脸识别研究与实现 [D]. 成都：电子科技大学，2016. (Wan Shining. Research and implementation of face recongnition based on convolution netural network [D]. Chengdu: University of Electronic Science and Technology of China, 2016. )

[18] 宋焕生，张向清，郑宝峰，等. 基于深度学习方法的复杂场景下车辆目标检测 [J]. 计算机应用研究, 2018, 35 (4): 1270-1273. (Song Huansheng, Zhang Xiangqing, Zheng Baofeng, *et al.* Vehicle detection based on deep learning in complex scene [J]. Application Research of Computers, 2018, 35 (4): 1270-1273. )

[19] 沈莉丽. 基于 Kinect 视觉识别的智能居家机器人系统 [J]. 组合机床与自动化加工技术, 2017 (12): 78-80, 84. (Shen Lili. Intelligent home robot control system based on the visual identity by kinect [J]. Modular Machine Tool & Automatic Manufacturing Technique, 2017 (12): 75-80, 84. )